音声言語処理の高度化に関する研究

伊藤　彰則
東北大学　大学院工学研究科

1．はじめに

コンピュータハードウェアの急激な高性能化と低価格化により、パーソナルコンピュータをはじめとする電子機器は広く普及している。これらの機器の性能は時間と共に向上しているが、それを利用するためのユーザインタフェースについては、1973 年に Xerox 社の ALTO ワークステーションに GUI(Graphical User Interface)が搭載されて以来、GUI を中心として構築されてきた。それ以降、OCR や手書き文字認識などの認識技術を利用したインタフェースが一部利用されてきているが、多くのユーザが日常的に利用するには至っていない。これに対し、音声認識やジェスチャ認識、顔画像認識等を利用した実世界指向インタフェースが近年盛んに研究されている。特に音声認識は、利用が容易であることや入力デバイスが安価であることなどから期待されている。

本研究の目的は、音声認識(特に連続音声の認識)を高精度化し、実用レベルの認識システムを構築することである。音声による発話内容を自動認識することで、音声によるコンピュータの操作はもちろん、講演の自動書き起こしや要約、テレビやラジオの自動索引付け、インターネット上の音声データの検索などさまざまな可能性が開ける。本研究では、このような応用のための基礎技術である「大語彙連続音声認識」の高精度化に関する検討を行う。大語彙連続音声認識には「音響モデル」「言語モデル」「辞書」「デコーダ」の４つの要素があるが、本研究ではこの中の言語モデルに焦点を当てる。言語モデルとは、認識する言語の持つ言語的な制約を表現したモデルであり、単語の系列がどれだけ言語的に妥当であるかを評価するものである。

本研究は大きく 3 つに分かれる。1 つは、N-gram 言語モデルのタスク適応に関する研究である[1][2]。N-gram 言語モデルは大量の言語データから作成される統計モデルであるが、作成の元になった言語データに強く依存する。そのため、例えば新聞記事から統計を取った N-gram の場合を利用した場合、人間同士の会話を認識しようとすると性能が著しく落ちる。本研究では、大量のデータ（例えば新聞記事データベース）に認識対象に近いデータを少量混入して統計を取ることで、認識性能を改善する技術を開発した。2 つ目は、確率文脈自由文法(SCFG)を大語彙音声認識に利用する研究である[3][4]。SCFG は以前から知られているモデルであり、音声認識に利用した研究も小数ながらあったが、モデル構築の計算量が膨大であったために広く利用されることはなかった。本研究では、日本語の特性(係り受け構造)を考慮して学習アルゴリズムを改良することにより、学習に必要な計算量を大幅に減らすことに成功し、初めて大語彙連続音声認識に SCFG を利用した。3 つ目は、上記の 2 つの研究に用いた言語モデル生成ツールをフリーソフトウェアとして整備し提供す

ることである[5][6][7]。

2．大語彙連続音声認識

まず、本研究の基礎となる大語彙連続音声認識について解説する。近年のほとんどの大語彙連続音声認識は、確率に基づく認識法の枠組みに基づいている。入力される音声信号を X、認識される単語系列を W とするとき、確率 $P(W|X)$ を最大にする $\hat{W}$ を選択する。ここで、Bayes の定理により

$$\begin{aligned}\hat{W} &= \underset{W}{argmax}\, P(W|X) \\ &= \underset{W}{argmax}\, \frac{P(X|W)P(W)}{\sum_{W} P(X|W)P(W)} \\ &= \underset{W}{argmax}\, P(X|W)P(W)\end{aligned}$$

すなわち、確率 $P(W|X)P(W)$ を最大にする単語列 $\hat{W}$ を認識結果とすればよいことになる。このとき、$P(W|X)$ を計算するモデルは「音響モデル」、$P(W)$ を計算するモデルは「言語モデル」と呼ばれる。また、W に含まれる単語集合を決定するのが辞書であり、定式を実際に計算して最適な認識結果を推定するプログラムはデコーダと呼ばれる。これを図１に示す。

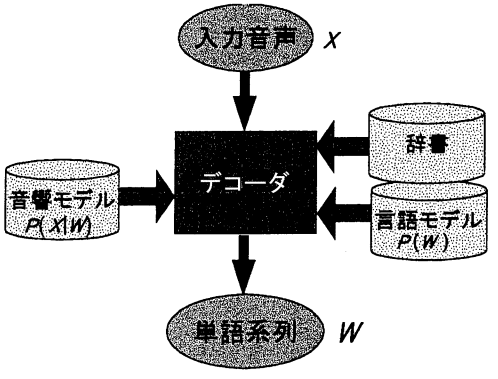

図１　大語彙音声認識システム

本研究では、大語彙音声認識の要素の中で言語モデルに着目して、従来のモデルをより高精度化する手法を検討する。

3．N-gram 言語モデルのタスク適応

第 1 の研究では、代表的な言語モデルである N-gram 言語モデルのタスク適応について検討した。N-gram モデルは現在最も広く利用されている言語モデルであり、ある単語の生

成確率を直前 N-1 単語による条件付確率として表現するモデルである。単語列 $w_1 w_2 \cdots w_n$ の生成確率は、N-gram モデルを使って次のように計算される。

$$P(w_1 \cdots w_n) = \prod_i P(w_i | w_{i-N+1} \cdots w_{i-1})$$

右辺の条件付確率を推定するため、大量のテキストデータから単語列の出現頻度の統計を取り、その出現頻度を計測する。例えば、N=3 の場合(trigram)には、条件付確率は次のように計算される。

$$P(w_i | w_{i-2} w_{i-1}) = \frac{N(w_{i-2} w_{i-1} w_i)}{N(w_{i-2} w_{i-1})}$$

ただし、$N(s)$ は文字列 s の出現頻度を表す。ここで推定される確率は、統計を取る元になったテキストデータに強く影響される。例えば、新聞記事データから統計を取って作成された N-gram モデルは、新聞記事に近い文体の文についてはよい推定値を与えるが、新聞記事と異なる文体の文については不当に低い確率を与える。そのため、そのような文を音声認識しようとすると、認識精度が低下する。原理的には、認識対象に近いテキストデータを大量に用意して統計を取ればよいのだが、実際にはそのようなデータを入手することが難しいことが多い。そこで、認識対象に近いが量が少ないデータと、認識対象には近くないが量が多いデータを混合して、可能な限りよい認識性能を与えるモデルを作成する。このとき、少量のデータでの文字列 s の出現頻度を $N'(s)$ とするとき、

$$P(w_i | w_{i-2} w_{i-1}) = \frac{N(w_{i-2} w_{i-1} w_i) + \gamma N'(w_{i-2} w_{i-1} w_i)}{N(w_{i-2} w_{i-1}) + \gamma N'(w_{i-2} w_{i-1})}$$

として確率を推定する。この式は Bernoulli 試行確率の Bayes 推定または MAP 推定の枠組みで捉えることができ、混合重み γ は事前確率分布として与えられるベータ分布のパラメータ α, β に基づいて決定することができる。この考え方に基づき、対話音声コーパスを用いたタスク適応実験を行い、さまざまな条件で適応させた場合の結果を求めた。その結果、単語誤り率が 24.7%から 16%まで減少し、誤り率で 35%の改善が得られた。

4．確率依存文法による連続音声認識のための言語モデル

第 2 の研究では、日本語の特性である「文節構造」と「係り受け制約」を SCFG（確率

文脈自由文法）に組み込む方法を提案した。SCFG は文脈自由文法に確率を付与したものであり、強力な言語モデルとして以前から知られているが、確率推定のための計算量が非常に多く、また確率の推定精度を上げることも難しいため、あまり用いられてこなかった。SCFG の確率の学習には Inside-Outside アルゴリズムと呼ばれる方法が使われる。この方法の計算量は、入力文の長さを L、非終端記号の数を N とするとき、$O(N^3 L^3)$ である。

これに対し、二つの方法によって計算量を大幅に削減することが可能となった。一つは、学習する文法の形を一般的な文脈自由文法から係り受け文法に制限することである。一般的な文脈自由文法の書き換え規則は

$$\alpha \rightarrow \beta\gamma \quad \alpha,\beta,\gamma \in N$$

のような形式であるが、係り受け文法の場合

$$\alpha \rightarrow \beta\alpha \quad \alpha,\beta \in N$$

とあらわすことができる。これにより、一つの規則を学習するのに考慮すべき非終端記号の種類が 3 個から 2 個に減るため、計算量が $O(N^3)$ から $O(N^2)$ に減少する。もう一つは、上の形式の規則の適用対象を「単語」から「文節」に変更することである。文節の中での単語連接規則（文節内文法）は正規文法で記述可能であるため、文節内正規文法の適用確率と文節間の文脈自由文法の確率を独立に推定することができる。この二つの方法を適用することによって、学習に要する計算量を従来の SCFG の 1/160 に削減することが可能になった。また、性能評価のために、新聞記事読み上げ音声の認識実験を行った。その結果、認識性能は従来用いられている言語モデルである trigram モデルと同等であり、trigram と SCFG を併用することによってさらに精度を高めることが可能になった。trigram の誤り率は 9.7%、SCFG 併用により 8.3%となり、誤り率は 14%削減された。

５．言語モデル作成ツールキットの開発

第 3 に、上記 2 つの研究成果を応用した言語モデル作成ツール palmkit を作成した。これは、上記の言語モデルを含む N-gram や SCFG を作成するためのツールキットであり、各種の N-gram 言語モデルだけでなく、タスク適応言語モデルなども簡単に作成することができる。本ソフトウェアは、情報処理学会連続音声認識コンソーシアムから「日本語ディクテーション基本ソフトウェア」の一部として配布されている。

〔参考文献〕

[1] 伊藤　彰則，好田　正紀：「N-gram 出現回数の混合によるタスク適応の性能解析」、電子情報通信学会誌(D-II), J83-D-II, 11, (2000), 2418-2427

[2] Akinori Ito, Hideyuki Saitoh, Masaharu Katoh and Masaki Kohda :"N-gram language model adaptation using small corpus for spoken dialog recognition", Proc.

European Conference on Speech Processing, (1997),2735-2738

[3] 堀　智織, 加藤　正治, 伊藤　彰則, 好田　正紀：「音声認識のための確率文脈自由文法に基づく言語モデルの構築と評価」、電子情報通信学会誌(D-II), J83-D-II, 11, (2000), 2407-2417

[4] Akinori Ito, Chiori Hori, Masaharu Katoh and Masaki Kohda:"Language modeling by stochastic dependency grammar for Japanese speech recognition", Proc. International Conference on Spoken Language Processing, (2000)

[5] 伊藤 彰則、好田　正紀：「単語およびクラス N-gram 作成のためのツールキット」、電子情報通信学会技術報告 SP2000-106 (2000), 67-72

[6] T. Kawahara, A. Lee, T. Kobayashi, K. Takeda, N. Minematsu, S. Sagayama, K. Itoh, A. Ito, M. Yamamoto, A. Yamada, T. Utsuro and K. Shikano: "Free Software Toolkit for Japanese large vocabulary continuous speech recognition", Proc. International Conference on Spoken Language Processing, (2000), 476-479

[7] http://palmkit.sourceforge.net/